



ISSN 2683-9067

[2025]

La EPH clara, accesible y mejorada

UN REPOSITORIO DIGITAL PARA ACCEDER, COMPRENDER Y
AMPLIAR LA INFORMACIÓN DE LA ENCUESTA PERMANENTE
DE HOGARES DE INDEC

La EPH clara, accesible y mejorada. Un repositorio digital para acceder, comprender y ampliar la información de la Encuesta Permanente de Hogares de INDEC

Informes del Observatorio UNR. N° 71

Informes Especiales N° 25

Marzo de 2025

Autor/es:

- Germán Adolfo Tessmer | german.tessmer@unr.edu.ar | <https://orcid.org/0000-0002-3827-7027>

Responsabilidad editorial:

- Germán Adolfo Tessmer

- ISSN (serie Informes del Observatorio UNR): 2683-9067

- ISSN (sub-serie Informes Especiales): 2683-9083

- Palabras Clave: EPH, Indicadores, Rstats, Manual estadístico

- Clasificación JEL: B41, C81

Esta obra está licenciada bajo la Licencia Creative Commons Atribución-NoComercial-Compartir Igual 4.0 Internacional.

Para ver una copia de esta licencia, visita <http://creativecommons.org/licenses/by-nc-sa/4.0/>.



Universidad Nacional de Rosario (UNR)

- Rector: Franco Bartolacci

- Vicerrector: Héctor Darío Masía



Observatorio Económico Social | UNR

<https://unr.edu.ar/observatoriounr/>

ISSN (página web) 2683-8923

Córdoba 1814 - Rosario, Argentina (S2000AXD)

+54 9 341 4802620 / Interno 147

observatorioeconsocial@unr.edu.ar

Palabras y problemas

Somos las historias que nos contamos, y los datos juegan un papel en eso.

El tema de nuestro tiempo en materia de conocimiento, es la capacidad de procesamiento. Pero no solo la que involucra el funcionamiento de una máquina sino, esencialmente, la capacidad que tenemos de procesar la información que recibimos del medio, incluida la que se deriva de la vida digital.

En perspectiva, nuestros cerebros siguen teniendo las características de un cazador-recolector, pero actualmente procesan información de múltiples identidades y medios digitales. Más emocionales que racionales, por diseño, nuestro problema consiste en discernir lo verosímil de lo que no.

La abundancia de información no es solo una cuestión cuantitativa. El volumen de información diaria al que está expuesta la mayoría de la población, plantea un cambio cualitativo en que las nociones consensuadas de verdad, ciencia, honestidad, justicia, o credibilidad. Si todo es discutible en un bajo nivel de complejidad, cualquier consenso se vuelve laxo. En el ruido aparecen los oportunistas, las fake news, los rage baits¹, los estafadores y los payasos.

Indudablemente, la información de calidad no puede llenar el vacío de sentido que asecha detrás de toda esta confusión. Pero sí puede contribuir a trazar límites entre aquello que busca reflejar cierta idea de realidad, contra aquello que sencillamente lo recicla en ficción. Un funcionario dice: “gracias a nuestras políticas de asistencia y contención social, más de la mitad de la población sigue contando con el respaldo del Estado para salir adelante, demostrando que nuestro gobierno no abandona a quienes más lo necesitan”. Mientras que otro tipo de discurso, menos florido y emotivo, puede afirmar que el dato registrado de pobreza fue de 38,9%.

Recibir información de calidad, implica que ésta ha recibido un procesamiento previo, una interpretación fundada, un procedimiento que sea replicable; una discusión honesta de ideas que pueda no coincidir con lo que pensamos, pero que tenga la potencia de ser contrastable. El problema con esto, es que la mayoría de la gente no tiene capacidad, formación, tiempo o voluntad de verificar; sobre todo si el tema en cuestión se encuentra lejano a su experiencia. Para esto recurrimos a mecanismos de confianza. Y de nuevo, en los márgenes pululan los oportunistas, y gurúes varios.

Las estadísticas sociales son un consenso. Los especialistas inventan una vara con la que miden algún fenómeno y quienes no son especialistas confían en eso. A mayor cantidad de intermediarios entre la medición y la noticia que involucra el dato relevado, mayor posibilidad de engaño. ¿Qué se puede hacer al respecto?

Si estudiar estadística no es una opción, entonces se pueden mejorar los mecanismos de confianza; y la universidad es uno de los ámbitos en los que definitivamente se puede realizar una contribución al respecto. La presentación de este trabajo va en línea con lo anterior. Presentamos el repositorio de datos [EPH – Observatorio](#) para toda la comunidad. De forma libre y gratuita, se presentan datos curados y ampliados de la [Encuesta Permanente de Hogares](#) relevada por [INDEC](#). Un pequeño aporte, en lo que consideramos, la dirección correcta.

¿Pero de qué trata la EPH? ¿Hay otras opciones? A continuación, una explicación.

¹ No existe aún una traducción al español para este término. Su traducción literal podría expresarse como “contenido provocador” o “anzuelo de odio”. En todo caso, el término se refiere al contenido diseñado deliberadamente para provocar indignación o reacciones emocionales fuertes. Este contenido suele encontrarse en redes sociales, sitios de noticias y foros. El objetivo es el de aumentar la interacción, aunque sea negativa, ya que la indignación genera mayor número de comentarios y reacciones y, por ende, mayor monetización del contenido.

La base EPH - Observatorio

La Encuesta Permanente de Hogares ([EPH](#)) es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos ([INDEC](#)), que permite conocer las características sociodemográficas y socioeconómicas de la población.

Indicadores tales como la tasa de actividad, de empleo y de desempleo son de suma importancia e interés para evaluar el desempeño de la economía de manera regular y consistente. Son fuente de noticias recurrente, y la base de esa información se obtiene gracias a las mediciones de la EPH.

El relevamiento es de carácter trimestral y cubre 31 aglomerados urbanos de la Argentina. Los microdatos de la encuesta son presentados por INDEC en dos bases de datos: individuos y hogares. Acompaña la presentación, el manual correspondiente para la interpretación de las bases mencionadas, además de documentos metodológicos.

Los datasets de la Encuesta Permanente de Hogares ([EPH - Observatorio](#)). Explicada, etiquetada y ampliada; consisten en una curadoría y ampliación de los datos originales realizada por el equipo del [Observatorio Económico Social | UNR](#) para ser utilizadas bajo el programa de código libre [R](#).

La información disponible en este repositorio, coincide con la última metodología de la base de microdatos propuesta por INDEC. Comienza en el cuarto trimestre de 2016 y contiene todos los trimestres siguientes hasta el último dato disponible.

Los datasets que se presentan en este repositorio representan un valor agregado con respecto a la fuente original (INDEC) en los siguientes aspectos:

- El uso intuitivo de la base de datos de EPH: Cada variable ha sido renombrada con un nombre no codificado, y etiquetada para que su significado pueda ser leído directamente desde la interfaz del programa. Lo mismo se ha hecho con los valores internos de cada variable. En vez de presentar números solo interpretables con el manual original de EPH, también se han etiquetado los significados de cada valor.
- Se han incorporado nuevas variables, que se pueden clasificar en tres grupos:
 - Producto del cálculo de dos o más variables dentro de la base original; como puede ser el caso del cálculo de Necesidades Básicas Insatisfechas (NBI).
 - Producto del cruce con datos externos a EPH; como puede ser las variables deflactadas de las distintas fuentes de ingresos.
 - Producto del desglose de clasificadores cerrados; como pueden ser las derivadas del Clasificador Nacional de Ocupaciones.

Estimamos que el público interesado en los datos procesados de la Encuesta EPH – Observatorio, puede ser:

- Investigadores y académicos: Especialistas en economía, sociología, demografía y ciencias sociales que podrían aprovechar la base de datos más accesible para realizar estudios sobre pobreza, empleo, desigualdad, distribución del ingreso, entre otros temas.
- Gobiernos y organismos públicos: Tanto a nivel nacional como provincial y local. Estos datos suelen ser utilizados para el diseño, implementación y monitoreo de políticas públicas relacionadas con el empleo, la educación, la salud y la vivienda. Variables como las Necesidades Básicas Insatisfechas (NBI) y los ingresos deflactados pueden proporcionar información valiosa para la planificación de políticas.
- Organizaciones no gubernamentales (ONGs): Las ONGs que trabajan en áreas como la reducción de la pobreza, la mejora de las condiciones laborales o la igualdad de género podrían utilizar estos datos para sustentar informes y acciones específicas.
- Consultoras y empresas privadas: Las consultoras dedicadas al análisis económico y social pueden emplear esta base de datos para realizar informes sectoriales y proveer recomendaciones a sus clientes en función de las tendencias y condiciones sociales que surgen de los datos.
- Periodistas y medios de comunicación: Con una base de datos más accesible e interpretativa, los periodistas podrían elaborar reportajes y análisis más precisos sobre la situación socioeconómica de distintas regiones de Argentina, brindando una mejor comprensión a la población en general.
- Organismos internacionales: Instituciones como el Banco Mundial, el BID, o la CEPAL, que monitorean y brindan apoyo para el desarrollo en América Latina, podrían reutilizar estos datos en sus estudios comparativos y análisis de políticas regionales.

En síntesis, el valor agregado de los datos más comprensibles y la incorporación de nuevas variables permiten que estos actores accedan a información clave de una manera más intuitiva y eficiente. Un pequeño aporte, en lo que consideramos, la dirección correcta.

Otros esfuerzos, diferentes accesos

La idea de mejorar la accesibilidad a los datos no es novedosa, ni ha sido única la forma de promoverlo. Dentro del entorno R, el programa de procesamiento de información estadística más comúnmente utilizado en la academia, existen otros esfuerzos similares, o incluso de mayor calidad, para favorecer el acceso a los datos de la población. El más conocido es el [paquete EPH](#)² que incluye los datos originales de INDEC organizados en una “colección de funciones, datos y documentación organizada en un formato estandarizado para ampliar las capacidades del lenguaje base”, realizada de forma colaborativa.

¿Cuáles son las ventajas de usar una u otra opción? La tabla siguiente muestra una comparación (indudablemente sesgada) de las ventajas y desventajas de utilizar uno u otro recurso.

² Cuyos autores, a inicios de 2025, son: Diego Kozlowski, Pablo Tiscornia, Guido Weksler, Germán Rosati, Natsumi Shokida y Carolina Pradier.

Tabla 1 | Tabla comparativa de atributos para bases de EPH

Base	Interpretabilidad	Curva de aprendizaje	Información ampliada	Peso de la base	Practicidad dentro del entorno R
INDEC (original)					
EPH Observatorio					
EPH R package					

Interpretabilidad

El gran volumen de usuarios de datos en nivel inicial utilizan hojas de cálculo (programas tales como Excel o Google Sheet) para interactuar con los datos en su actividad. Asimismo, no todas las organizaciones disponen de licencias de softwares de mayor complejidad por una simple razón de escala. Pequeñas organizaciones con el núcleo de actividad especializado en otra área de negocio, sencillamente utilizan herramientas digitales de oficina; más genéricas y que no requieren mayor inversión en formación de talento. Y, por supuesto, hay que sumar a lo anterior la inercia de cada organización, sobre todo las grandes sin mayor base tecnológica, incluidas las estatales.

La base original de EPH presenta información codificada que hace difícil su interpretación directa para usuarios como los descriptos anteriormente. Por supuesto, la presentación de INDEC cumple con los requisitos de normalización de la información, que minimizan la redundancia de registros, pero a costa de sumar complejidad en su interpretación.

Gráfico 1 | Ejemplo de la visualización de la base de datos de EPH INDEC

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	MAS_500	AGLOMERADO
1	TQRMNORUJHLMRCDEIAD00861800	2024	2	1	2	1	1 S		32
2	TQRMNORPRHKLMLUCDEIAD00858515	2024	2	1	1	1	1 S		32
3	TQRMNORPRHKLMLUCDEIAD00858515	2024	2	1	2	1	1 S		32
4	TQRMNORPRHKLMLUCDEIAD00858515	2024	2	1	3	1	1 S		32
5	TQRMNOSUHLMLUCDEIAD00858728	2024	2	1	1	1	1 S		32
6	TQRMNOSUHLMLUCDEIAD00858728	2024	2	1	2	1	1 S		32
7	TQRMNOQWVHKMNLCEIAD00858516	2024	2	1	1	1	1 S		32
8	TQRMNOQWVHKMNLCEIAD00858516	2024	2	1	2	1	1 S		32
9	TQRMNOQWVHKMNLCEIAD00858516	2024	2	1	3	1	1 S		32
10	TQRMNOQWVHKMNLCEIAD00858516	2024	2	1	4	1	1 S		32

La base EPH-Observatorio provee una solución en ese sentido, al presentar un uso más intuitivo de la información a nivel usuario. Cada atributo (columna) de la base presenta un nombre no codificado y una breve explicación legible directamente en la base de datos. Asimismo, las bases vienen acompañadas de su [correspondiente manual](#), donde se explican las transformaciones aplicadas a cada atributo.

Gráfico 2 | Ejemplo de la visualización de la base de datos de EPH - Observatorio

	codusu Código viviendas	nro_hogar Código hogares	id_hogar Identificador de hogar	realizada Entrevista realizada	año Año	trimestre Trimestre	region Código región	mas_500 Tamaño aglomerado	aglomerado Código Aglomerado
1	TQRMNORUJLMRCDEIAD00861800	1	TQRMNORUJLMRCDEIAD00861800+1	Si	2024	2	GBA	Si	Ciudad de BsAs
2	TQRMNORPRHKLMLUCDEIAD00858515	1	TQRMNORPRHKLMLUCDEIAD00858515+1	Si	2024	2	GBA	Si	Ciudad de BsAs
3	TQRMNORPRHKLMLUCDEIAD00858515	1	TQRMNORPRHKLMLUCDEIAD00858515+1	Si	2024	2	GBA	Si	Ciudad de BsAs
4	TQRMNORPRHKLMLUCDEIAD00858515	1	TQRMNORPRHKLMLUCDEIAD00858515+1	Si	2024	2	GBA	Si	Ciudad de BsAs
5	TQRMNOSUHLMLUCDEIAD00858728	1	TQRMNOSUHLMLUCDEIAD00858728+1	Si	2024	2	GBA	Si	Ciudad de BsAs
6	TQRMNOSUHLMLUCDEIAD00858728	1	TQRMNOSUHLMLUCDEIAD00858728+1	Si	2024	2	GBA	Si	Ciudad de BsAs
7	TQRMNOQWVHKMNLCEIAD00858516	1	TQRMNOQWVHKMNLCEIAD00858516+1	Si	2024	2	GBA	Si	Ciudad de BsAs
8	TQRMNOQWVHKMNLCEIAD00858516	1	TQRMNOQWVHKMNLCEIAD00858516+1	Si	2024	2	GBA	Si	Ciudad de BsAs
9	TQRMNOQWVHKMNLCEIAD00858516	1	TQRMNOQWVHKMNLCEIAD00858516+1	Si	2024	2	GBA	Si	Ciudad de BsAs
10	TQRMNOQWVHKMNLCEIAD00858516	1	TQRMNOQWVHKMNLCEIAD00858516+1	Si	2024	2	GBA	Si	Ciudad de BsAs

Para mantener la coherencia entre las variables originales provistas por INDEC y las mismas renombradas en la base EPH – Observatorio, el manual provee una suerte de fichaje de cada variable, como la que se muestra en el siguiente ejemplo:

Gráfico 3 | Ficha explicativa de la variable busco_trabajo_anio (PP02H) tomada de ejemplo del Manual EPH - Observatorio

busco_trabajo_anio 1.	
2. Variable Original: PP02H	3. Tipo de variable: Factor
- Significado: En los últimos 12 meses, el miembro del hogar, ¿buscó trabajo en algún momento? 4. - Significado resumido: Buscó trabajo en el último año 5.	
6. Base: Individuos	7. Categoría: Características de los miembros del hogar
Valores	Significado de la respuesta en la variable original
Si 9.	1 = Si 8.
No	2 = No

De donde:

- Es el nombre de la variable tal como aparece en la base de datos. Es decir, el nombre recodificado de la variable.
- Es el nombre original de la variable, tal como aparece en las bases de individuos y hogar de EPH.
- El tipo de variable, muestra cómo ha sido condicionada la variable en R.
- El significado de la variable según el manual de EPH, pero reinterpretado según la categoría a la que pertenece la variable. Con esto se ha querido que la lectura del significado sea completa variable a variable.
- El significado resumido, es la etiqueta de la variable que figura en la base de datos. Cuando en una variable figura solo Significado, significa que el punto 4 y 5 coinciden, y que se considera que no se necesita mayor explicación.
- Base, hace referencia a la base de datos a la que pertenece la variable. Contiene:
 - Individuos: variables que solo aparecen en la base de individuos.
 - Hogar: variables que solo aparecen en la base hogar.
 - Identificación: variables que aparecen en la base individuos y hogar original, o que cumplen funciones de identificación

- d. Hogar e individuos: variables que han sido creadas del cruce de información entre las variables hogar e individuos.
7. Categoría, es la agrupación a la que pertenece la variable en el contexto del manual original de EPH o bien, la agrupación creada a los efectos de clasificar de mejor manera a las variables creadas.
8. El significado de la respuesta en la variable original, muestra los valores y significados en las bases originales de EPH.
9. Valores, son los valores o etiquetas que figuran en las bases EPH – Observatorio.

Finalmente, el paquete EPH, el desarrollo más sofisticado de los tres que se están comparando, también presenta un etiquetado de variables opcional. En este caso, la estrategia de los autores ha sido la de preservar el nombre original de las variables, pero mostrando un etiquetado para cada una de estas. El resultado, es el que se muestra en el gráfico siguiente:

Gráfico 4 | Ejemplo de la visualización de la base de datos del paquete EPH en R

	CODUSU Codigo para distinguir VIVIENDAS, permite aparearlas con Hogare...	ANO4 Año de relevamiento	TRIMESTRE Ventana de Observacion	NRO_HOGAR Codigo para distinguir HOGARES, permite aparearlos con Personas	COMPONENTE Numero de componente: N° de orden que se asigna a las per...
1	TQRMNORUUHJLMRCDEIAD00861800	2024	2	1	2
2	TQRMNORPRHKLUMUCDEIAD00858515	2024	2	1	1
3	TQRMNORPRHKLUMUCDEIAD00858515	2024	2	1	2
4	TQRMNORPRHKLUMUCDEIAD00858515	2024	2	1	3
5	TQRMNOSSUHKLMUCDEIAD00858728	2024	2	1	1
6	TQRMNOSSUHKLMUCDEIAD00858728	2024	2	1	2
7	TQRMNOQWVHKMNLCDIAD00858516	2024	2	1	1
8	TQRMNOQWVHKMNLCDIAD00858516	2024	2	1	2
9	TQRMNOQWVHKMNLCDIAD00858516	2024	2	1	3
10	TQRMNOQWVHKMNLCDIAD00858516	2024	2	1	4

¿Por qué consideramos que nuestra base es la que mejor interpretabilidad tiene? Primero, porque la lectura es directa incluso a costa de sacrificar los nombres originales de las variables (por supuesto, esta es una decisión polémica, y depende del grado de uso que se tenga de la base EPH), y también, por la forma en que se encuentra organizada la información respaldatoria de cada variable. El paquete EPH es la segunda mejor opción, quizás porque requiere pasos adicionales para obtener el etiquetado, que por lo demás no son complicados de implementar; pero que también son más extensos. Finalmente, la información provista por INDEC es la que menor interpretabilidad directa tiene, aunque hay que admitir que recientemente han mejorado el manual respaldatorio de variables, haciendo más amena su lectura.

Curva de aprendizaje

Familiarizarse con la EPH lleva un tiempo de aprendizaje sobre dos aspectos: la información en sí, y en cómo procesarla. Tal como están presentadas las bases en las tres opciones, este punto se encuentra íntimamente entrelazado con el anterior: una exposición amena de la información, favorece su interpretación, y a través de ésta, su aprendizaje.

Por otra parte, la utilización de cada versión de la base de datos de la EPH requiere del uso de alguna herramienta para su procesamiento. En este aspecto, la información original provista por INDEC es más libre, dado que los datos se

presentan en formato .csv, que es uno de los más sencillos de cargar en prácticamente cualquier programa de procesamiento estadístico. En cambio, las otras dos versiones requieren el uso de R como paso inicial.

A diferencia de Python, “el” software de código libre para análisis de datos y machine learning, R es un programa que actualmente ha encontrado su nicho entre usuarios pertenecientes al entorno académico, estadísticos, profesionales del sector público y ONGs, actuarios, y especialistas en visualización de datos y en reporting automatizado. El mismo tiene una curva de aprendizaje de pendiente alta al comienzo, la que puede llegar a apabullar a potenciales nuevos usuarios. En todo caso, resulta natural que información como la provista por INDEC no sea presentada en este programa.

En el caso de las bases EPH – Observatorio, con solo unos pocos pasos, se pueden exportar las bases al formato de hoja de cálculo, es decir, al entorno más comúnmente usado por el público general. Lo mismo aplica para las bases del paquete EPH, pero se requieren unos pocos pasos adicionales para lograr el mismo cometido. Por este motivo, es que en la **Tabla 1** se asigna a cada alternativa en ese ranking.

Información ampliada

¿A qué se denomina información ampliada en este contexto? A la elaboración de información y puesta a disponibilidad de información adicional basada en los datos originales de la EPH. En este sentido, el paquete eph, por lejos, es el mejor rankeado dado que incorpora dentro del entorno R la posibilidad de: iterar variables, construir pools de datos de panel y realizar cálculos de línea de pobreza, entre otros.

Por lógica, le sigue en el ranking las bases EPH-Observatorio que incorporan variables de ingresos deflactadas por región, y cálculos sencillos de necesidades básicas insatisfechas. Finalmente, la información procesada de INDEC no se encuentra en las bases, sino en la documentación de publicación periódica del instituto.

Peso de la base

Este aspecto es el punto débil de la estrategia propuesta en las bases EPH-Observatorio, dado que el costo de la accesibilidad se paga con mayor espacio de almacenamiento y en tiempo de procesamiento. Estas bases incumplen sistemáticamente con los protocolos de normalización de información, contrario a como se presentan en INDEC. De nuevo, y cómo se explicó en el primer punto, todo depende del público al que se busca servir. En este sentido, la estrategia del paquete eph se encuentra bien balanceada, dado el etiquetado de la base es opcional en vez de ser compulsivo.

Practicidad dentro del entorno R

El paquete eph ocupa sin lugar a dudas el podio en términos de practicidad para el manejo de R. Las herramientas incorporadas en forma de funciones ejecutables para el cálculo de información ampliada lo convierten en la mejor opción para usuarios avanzados del entorno. Ocupa el segundo lugar nuestras bases de datos, dado que la carga es directa y diseñada para este programa.

Para cerrar

Uno de los mayores problemas a la hora de lidiar con la EPH, no mencionado aun en el documento, es la difícil replicabilidad de los datos que se proveen en los informes regulares del INDEC. Si bien las metodologías se encuentran ampliamente reportadas, al ejecutarlas aparecen pequeñas diferencias que indudablemente requieren atención por parte de los usuarios.

Para quienes tomen el desafío de ir un paso más allá y tratar de arribar a sus propias conclusiones, las opciones alternativas aquí mostradas facilitan ese camino. Además, hay más de una, con sus pros y sus contras. En el caso de las bases EPH-Observatorio un pequeño aporte, en lo que consideramos, la dirección correcta.